The Estimation of Mutual Funds' Holdings Based on Constrained Kalman Filter

Ziao Zhang Zongliang Han

December 14, 2023

1 Introduction

Mutual funds, as significant participants in the capital market, attract widespread attention to their investment trading behavior. The changes in their holdings can reflect institutional investors' perspectives on the market, offering valuable guidance for constructing trading strategies. However, holding information of mutual funds, that is the proportion of investment in each industry, is only disclosed in the annual and semi-annual reports. Therefore, alternative methods are needed to obtain high-frequency estimates of fund holdings. For example, a certain mutual fund disclosed in its annual report that it invested 10% of its funds in the steel industry, 40% in the electronics industry, 30% in the pharmaceutical industry, 20% in the environmental protection industry, and the remaining 10% in government bonds. The problem is, after three months, we want to know the changes in this fund's investment portfolio.

2 data

We adopt Shenwan Hongyuan's industry classification, which divides the entire market into 31 industries, including steel, pharmaceuticals, coal, banking, automobiles and others. Mutual funds can allocate their investment portfolios among these 31 industries and government bonds(To simplify, we consider government bonds as an industry too).
And, we use industry index to track and measure the performance of a specific industry within an economy. It reflects the overall changes in the prices of a group of stocks or assets belonging to that particular industry. Industry indices are constructed by compiling the prices or other relevant indicators of representative companies within that industry. They are used by investors to assess the health, trends, and investment opportunities within specific sectors. We can easily obtain the daily return series of these indices from December 31st, 2022, to June 30th, 2023
We have selected 30 funds for estimation purposes. For each fund, we have daily return series of the fund itself, also from December 31st, 2022, to June 30th, 2023. Additionally, we have the disclosed holdings from the semi-annual report on June 30th, 2023.
Based on these data, our objective is to estimate the industry holdings of the funds on June 30th, 2023. Furthermore, we intend to validate these estimations using the disclosed holdings from the semi-annual report.

3 traditional method

The traditional estimation method is to fit the constrained lasso regression with the fund's return with industry indices' returns. The coefficient at each industry index represents the estimated holding.
We know that Lasso regression can enable some regression coefficients to become zero. From this point, it is easy to understand why we chose lasso regression. Very few fund managers have the ability to allocate funds across all industries.
However, when we apply lasso regression, what we really do is to use data from a certain period in the past for estimation, logically we can only obtain the average holdings over that period, rather than the current holdings at a specific time.

4 Kalman filter

We find that we can treat the industry holdings of the fund as hidden states and the fund's returns as observations. Then, we model the fund returns using a state-space model. Efficiently, with the Kalman filter, we can identify changes in the fund holdings.
The model is

$$\begin{cases} w_k = w_{k-1} + e_k & e_k \sim N(0, Q) \\ R_k = r_k w_k + \eta_k & \eta_k \sim N(0, R) \end{cases}$$

w_k is the hidden state, which represents the holdings on time k. R_k is the observation, which represents the fund's return on time k. r_k is the observation matrix, which represents the return of 32 industry indices.
As it mentions above, w_k is the holdings of fund on time k, so we must impose some constrains on it, that is

$$\begin{cases} w_k \geq 0 \\ \sum_{i=1}^{32} w_{k,i} = 1 \end{cases}$$

We choose two different methods to solve the constrained problem. The first is to solve the constrained kalman filter directly and we call it constrained kalman, that is

$$\begin{cases} w_k = w_{k-1} + e_k & e_k \sim N(0, Q) \\ R_k = r_k w_k + \eta_k & \eta_k \sim N(0, R) \\ s.t. & w_k \geq 0 \\ & \sum_{i=1}^{32} w_{k,i} = 1 \end{cases}$$

The second method is to adjust w_k computed from the unconstrained model on time k. In detail, it contains two steps. The first step is to set all weights less than 0 to 0. Then we normalize it. We call it adjusted kalman.

5 error comparison

For 32 industries and 30 funds, we can derive the result with the form below, where e_{ij} is the absolute error.

	industry 1	industry 2	...	industry32	rowsum
fund 1	$e_{1,1}$	$e_{1,2}$	...	$e_{1,32}$	$\sum_i e_{1,i}$
fund 2	$e_{2,1}$	$e_{2,2}$	...	$e_{2,32}$	$\sum_i e_{2,i}$
...	...	...	...	...	
fund 30	$e_{30,1}$	$e_{1,2}$	...	$e_{30,32}$	$\sum_i e_{30,i}$
colsum	$\sum_i e_{i,1}$	$\sum_i e_{i,2}$	...	$\sum_i e_{3,32}$	

To access the accuracy of our estimation, we choose two ways to calculate the mean error of these three method.
First we calculate the mean of the rowsum(see table above) and we define it as fund error. It shows the error on each fund
The fund error is

$$\frac{1}{30} \sum_k \sum_i e_{k,i}$$

The result shows below

method	error 2
Lasso regression	38%
Constrained Kalman filter 1	0.037%
Adjusted Kalman filter 2	30.38%

We can see that the constrained kalman has a very tiny error compared to Lasso regression and Adjusted Kalman. In fact, due to the significant size of the funds we've selected, the fund managers have made relatively small adjustments in portfolio reallocation. Both of the lasso regression and adjusted Kalman filter make great adjustments in portfolio reallocation.
Besides, we use the least square estimation to solve the constrained Kalman, it gives more trust on the prior rather than the observations compared to the MLE. It can also explain why the constrained Kalman has an amazing result.
For the adjusted Kalman, we make a strong change of the state at each step after we get the unconstrained model. However, we do not make any changes on the convariance matrix. And, the introduced constraints during the iterations of the Kalman filter did not provide effective exogenous information; instead, it disrupted the optimal characteristics of the original iteration process.
Then we calculate the mean of the colsum(see table above) and we define it as industry error. It shows the error on each industry
The industry error is

$$\frac{1}{32}\sum_{k}\sum_{i} e_{i,k}$$

The result shows below

method	error 2
Lasso regression	113%
Constrained Kalman filter 1	20.6%
Adjusted Kalman filter 2	86.3%

Although industry error is much larger than the fund error, we can also see the constrained kalman is relatively correct.
Reconsider the features of these 30 funds, we found that most of the funds have relatively diversified holdings, which means the fund managers made investments across many industries. We know that lasso regression will enable some regression coefficients to become zero. And we enforce the state to become zero under the adjusted kalman filter. So the estimated holdings are expected to be concentrated in several specific industries if we use lasso regression and adjusted kalman filter. That's why they have a really huge error.
Figure 1 shows the error on different industries. It also shows that adjusted kalman and lasso regression have huge error on some specific industries such as power equipment and government bonds.

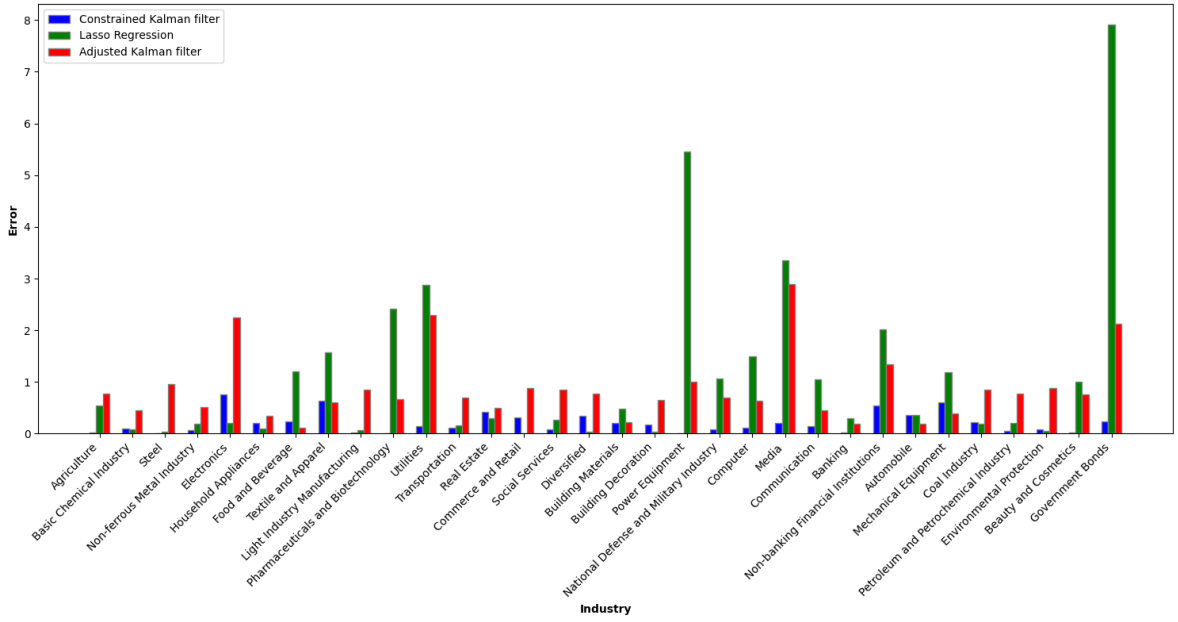

Figure 1: industry error

Figure 2 shows the error on different funds and we can see the constrained kalman filter is effective on every funds.

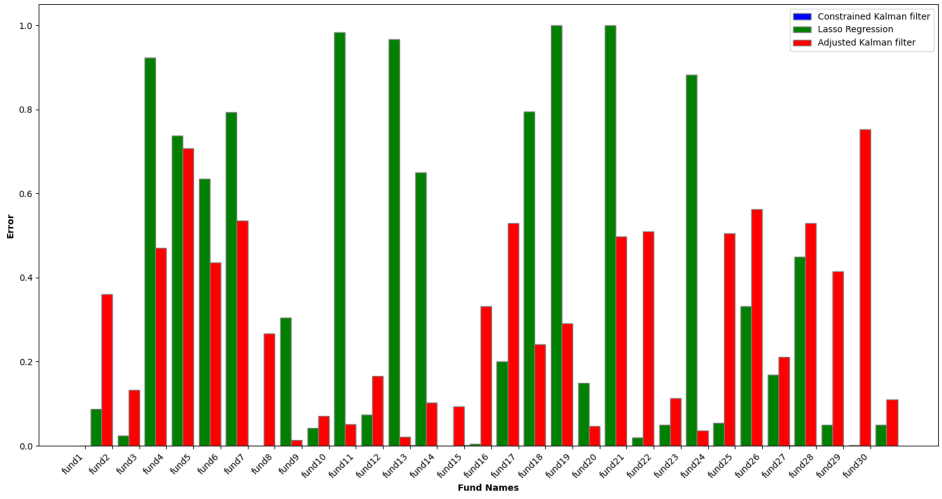

Figure 2: fund error

For a specific fund, figure 3 shows the real holdings and estimated holdings. We can see the constrained kalman has a great performance.

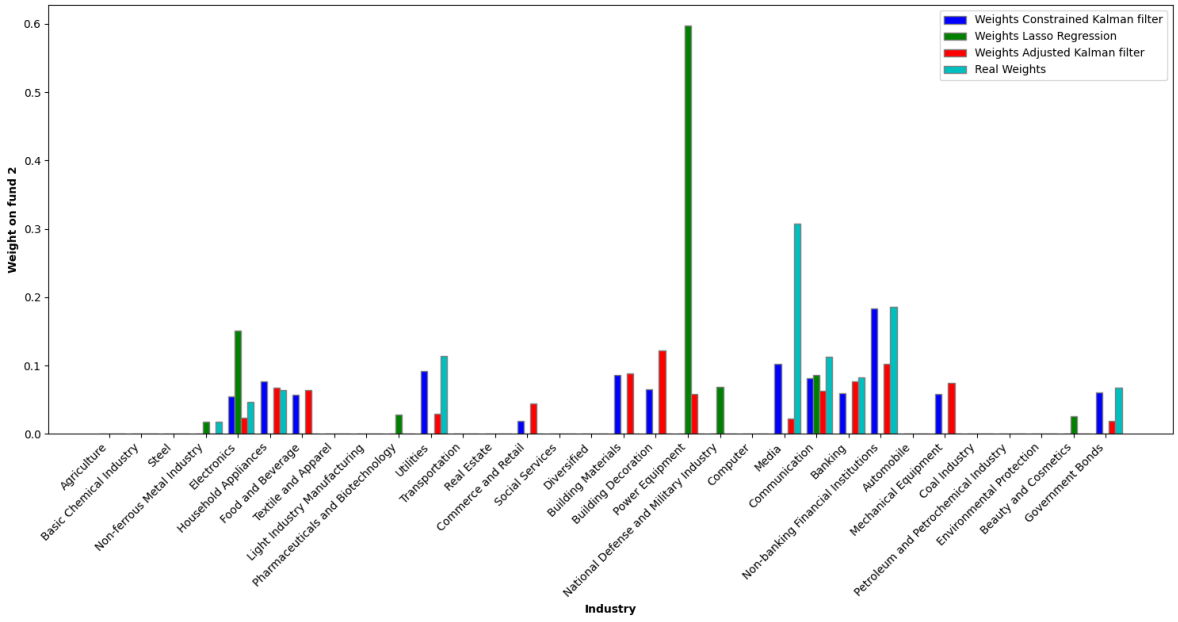

Figure 3: fund 2

6 conclusion

We treat the industry holdings of the fund as hidden states and the fund's returns as observations and establish a state space model on the fund holdings. We take two methods to derive the kalman filter. Both of the constrained kalman and adjusted kalman are more effective than the traditional lasso regression. Among which, constrained kalman shows an excellent result on every fund and this method can give a valuable guidance on the investment strategies. Notice that we only select 30 funds to get the conclusion, perhaps it can not represent the majority. These 30 funds also show some similar futures. So we need to expand our date set in the subsequent studies to get a more rigorous conclusion.